ECHO Ingest Operations: Quarterly Report #1
09/13/2003 –12/19/2003

DRAFT FOR ECS DAAC REVIEW

Contents

Introduction

This document provides information on the acquisition and processing of metadata files received for ingest into the ECHO operational system (api.echo.eos.nasa.gov) following the release of version 5.0 on 09/10/2003.

With the release of version 5.0, a complete refresh of the operational system metadata catalog was performed for four of the six data partners who have been actively participating in ECHO since its operational implementation in November 2002. The metadata refresh was performed at the request of the NASA EOS Core System (ECS) Distributed Active Archive Center (DAAC) data partners: GES DAAC, LaRC DAAC, LP DAAC, and NSIDC DAAC.

Since metadata records from the four ECS DAACs represent the majority of metadata required to enable ECHO as a viable alternative to existing infrastructure, i.e., support the new EOS Data Gateway (EDG) system being developed as an ECHO client application, the ECHO Operations Group (ECHO Ops) considers the refresh at version 5.0 a logical mark for the beginning of ECHO ingest analysis. As such, we are presenting information from the 14-week period that spans 09/13 – 12/19/2003 in this first quarterly report of ECHO ingest operations.

In addition to providing metrics on ECHO ingest for project stakeholders and external participants, the information provided in this document was collected and analyzed in order to help us:

- Identify critical short and long term issues for ECHO operations management and continued system development;
- Identify requirements and specifications for an ingest management and accounting infrastructure;
- Adjust ECHO Ops plans and procedures as needed to meet goals set for the ECS DAAC historical load ingest.

Part I: Ingest Performance

The ECHO ingest process currently consists of the following steps:

1) Generation of metadata files by data partners;
2) Transport of metadata files to the ECHO ingest system via ftp-push by data partners;
3) Review and pre-processing (if necessary) of metadata files by ECHO Ops;
4) Staging of metadata files for ingest by ECHO Ops;
5) Processing of metadata files by the ECHO system;
6) Review of ingest results by ECHO Ops and data partners.

Although ECHO was designed to automatically process metadata files received from data partners (by polling the partner ingest ftp directories at routine intervals), ECHO Ops introduced steps 3 and 4 at the start of version 5.0 operations in order to address several key issues that were identified in early ingest operations:

- Because of the serial nature of the current ingest system[1], one data partner can dominate ingest for an extended period of time by sending a large number of files.

- Data Partners do not necessarily want metadata files to be ingested right away or in the order that they are transmitted;

- There are strategic reasons for manually controlling the order of ingest processing, such as prioritizing ingest to meet client application needs;

- There are technical reasons for manually controlling ingest processing, such as short-term needs for pre-processing metadata files to prevent undesired transactions (e.g. insertion of granule records with invalid spatial parameters that were identified but not being excluded from ingest in version 5.0).

In order to derive metrics and examine progress made during the period under review, ECHO Ops collected and organized information for the factors that affect ECHO catalog holdings: metadata flow to ECHO; metadata available and processed; and ingest rates.

1. Metadata Flow to ECHO

Information on metadata acquisitions was collected by performing a retrospective scan of files stored on the ingest system. Before, during, and after ingest processing, the metadata files are distributed among directories organized by ProviderId and their stage in the ingest process. By scanning directories, extracting file system file information, parsing the files themselves for relevant parameters (XML tags), we were able to create a database (table IngestFile) containing information on all ingest files received during the period. Some additional manual processing was needed in order to classify the files as historical load or forward processing. From this database, we were able to aggregate and analyze the ingest file information as needed. **Table 1** provides a summary of the metadata acquisitions (collection, granule, and browse XML files) from all data partners by week.

[Please note: Data from Weeks 01 and 02 must be confirmed from files in the process of being restored from backup. Table 1 and subsequent tables/figures may be updated in preparation of the final version of this report.]

Week	Dates	Files	Records	Megabytes
01	09/13/03 – 09/19/03	35	124641	1636
02	09/20/03 – 09/26/03	2	131	1
03	09/27/03 – 10/03/03	284	447682	2353
04	10/04/03 – 10/10/03	231	965613	8280
05	10/11/03 – 10/17/03	220	1687325	6956
06	10/18/03 – 10/24/03	962	1309279	7162
07	10/25/03 – 10/31/03	236	1511490	11999

[1] In its current implementation, the ECHO ingest system looks for files by provider entity (e.g. GSFCECS), and it will process all files available for the current provider before moving on to another provider.

08	11/01/03 – 11/07/03	210	2180917	12413
09	11/08/03 – 11/14/03	219	1255931	7804
10	11/15/03 – 11/21/03	1021	1964092	8380
11	11/22/03 – 11/28/03	2213	2639123	13011
12	11/29/03 – 12/05/03	4128	2720656	12692
13	12/06/03 – 12/12/03	3561	1375884	3871
14	12/13/03 – 12/19/03	332	281712	1500
Total	**09/13/03 – 12/19/03**	**13654**	**18464476**	**98058**

Table 1. Metadata received for ingest in ECHO operational system during the reporting period.

In addition to the XML ingest files, we received more than one million binary browse files (> 192 GB) during the period.

Figure 1a provides a summary of total records received by Provider by week. In addition to the records shown in the figure, we received 6703 records for metadata replacement from the ORNL DAAC in 19 files on one day in week 4.

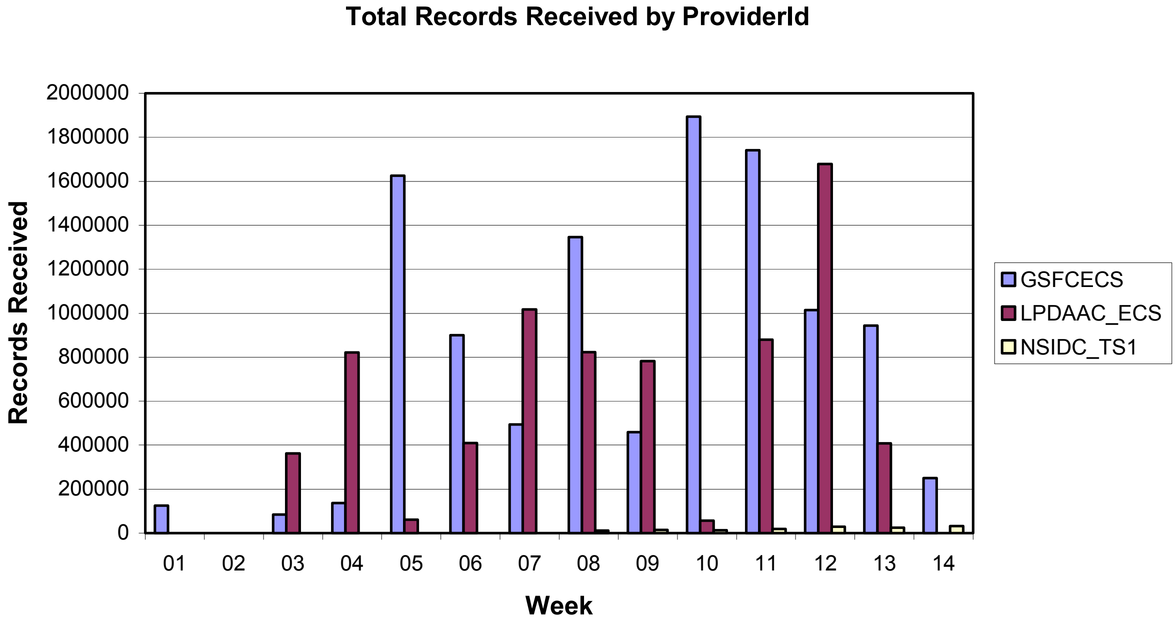

Figure 1a. Records received for ECHO operational system ingest by Provider ID.

Figure 1b provides a summary of forward processing records received by Provider by week.

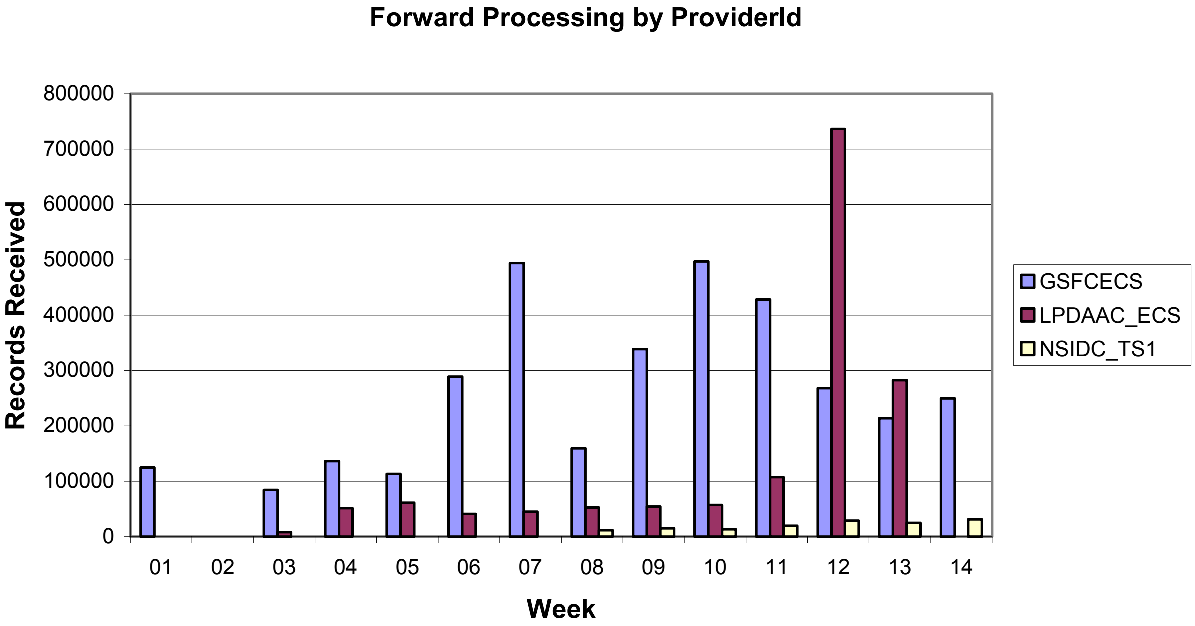

Figure 1b. Forward processing records received for ECHO operational system by Provider.

2. Metadata Available and Processed

Metadata files transferred to the ECHO operational system are not necessarily available for immediate ingest. Sometimes data partners request that the metadata files be held for later ingest. Other times there are technical issues with the files that prevent ingest.

In order to examine progress in metadata processing, we needed to determine the first date on which the files received were available for processing. Ideally, this would be captured from the filesystem date/time recorded immediately following file transfer. Unfortunately, this date/time was not captured, and subsequent activity on the ingest system altered many of the original file date/time stamps. In order to provide a consistent measure of the initial receipt date/time, we used the data partner creation date extracted from the ingest file naming scheme for most of the files analyzed. This field was inserted into the IngestFile table as ProviderDate. For example, in the naming convention used by the Bulk Metadata Generation Tool (BMGT), we extracted ProviderDate from the character string that is located to the left of the ".XML" in the file name (e.g. for file "EDCGASTT200328020032810101.20031008003826.XML", ProviderDate is 20031008003826 or 10/08/2003 00:38:26).

The date available for processing was then determined by reviewing our data partner issue and action item records[2]. Where a technical issue or data partner request affected file availability for processing, the DateAvailable field in the IngestFile table was manually set to the appropriate value; otherwise, the DateAvailable was set to the value extracted for ProviderDate.

Similar to metadata acquisition, information on metadata processing was collected by performing a retrospective scan of ingest log files stored on the operational system. To extract information, we parsed the ingest log files to extract the start date/time for the ingest job, names of the metadata files processed, and a variety of transaction metrics (see *Part I: Section 4*). From the extracted information, we created a database table (LogFile) that contains records from all log files created during the period and a table (Processed) that associates metadata ingest files with the appropriate ingest jobs recorded in the LogFile table. From this database, we were able to aggregate and analyze the ingest job information as needed.

Table 2 provides a summary of metadata available and processed by week.

Week	Cum. Records Available	Cum. Records Processed	Available Processed (%)	Staged More Than Once (%)	Comments
01	TBD	TBD	TBD	TBD	GSFCECS granule ingest began several days prior to the start of the period (09/09).
02	TBD	TBD	TBD	TBD	
03	363198	363198	100.0%	0.0%	Ingest operations suspended 10/01-02 due to a planned system upgrade and an unexpected HW failure; LPDAAC_ECS ingest began on 10/03.
04	396604	396604	100.0%	0.0%	
05	1243560	1243560	100.0%	20.3%	
06	2021572	2020969	100.0%	64.0%	
07	2111111	2110508	100.0%	61.5%	
08	3374685	2674667	79.3%	54.1%	Browse ingest was suspended on 11/06 due to system HW problems; NSIDC_TS1 ingest began on 11/07.
09	3616536	2916518	80.6%	60.7%	No browse ingest performed this week.
10	8443925	5332391	63.2%	30.5%	No browse ingest performed this week; LPDAAC gave approval for ingest of Terra and Aqua MODIS collections, resulting in a 133% increase in records available.
11	8883429	5471468	61.6%	29.0%	No browse ingest performed this week.
12	9647956	5569748	57.7%	26.9%	No browse ingest performed this week; version 5.0.1 transition to ops began 12/02.
13	10217323	5662139	55.4%	25.4%	Browse ingest restarted on 12/10; version 5.0.1 ingest began on 12/10.
14	15171678	6010173	39.6%	17.3%	System problems and ultimate HW failure impacted ingest during 12/16-17; GSFCECS granule files previously held (due to invalid spatial parameters that could not be handled properly in version 5.0) became available for ingest this week,

[2] Since mid-November (2003), these records have been provided to the DAACs and other ETC participants on a weekly basis in the ECHO Ops Weekly Status Report distribution package.

					creating a 48% increase in records available.

Table 2. Metadata processed in ECHO operational system during Quarter #1.

Figure 2 provides a summary of total records processed for Provider by week.

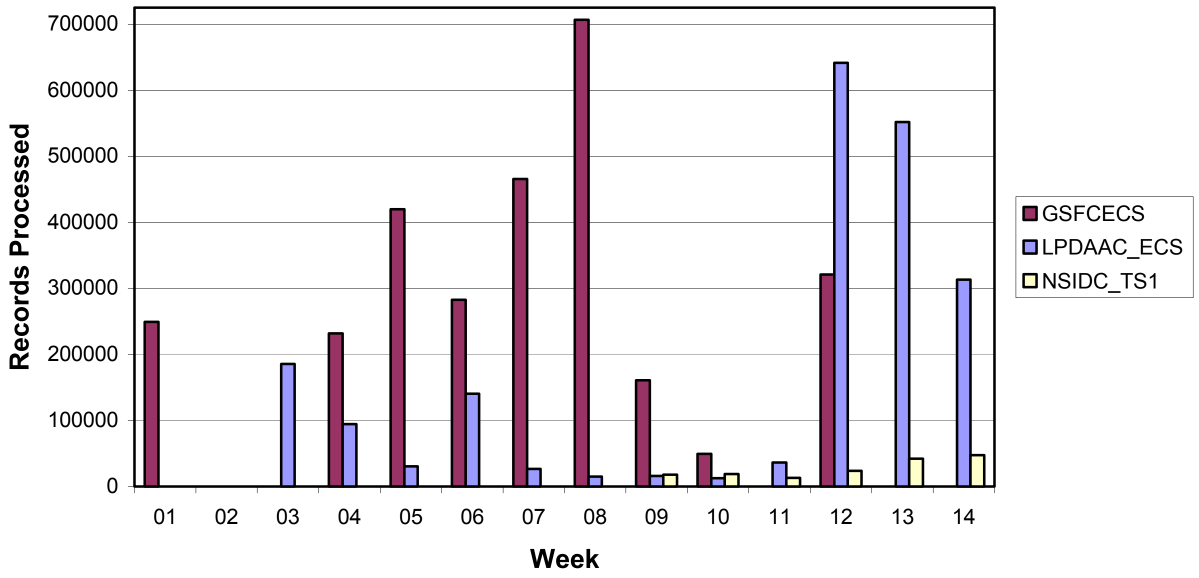

Figure 2. Records processed in ECHO operational system during Quarter #1 by Provider by week.

3. Accumulated Backlog

Figure 3 provides a summary of the records received but not processed during the quarter. The 12,447,397 unprocessed records shown in the figure are distributed as:

• GSFCECS granules:	5,564,801
• GSFCECS browse:	2,316,172
• LPDAAC_ECS granules:	4,546,937
• LPDAAC_ECS browse:	19,487

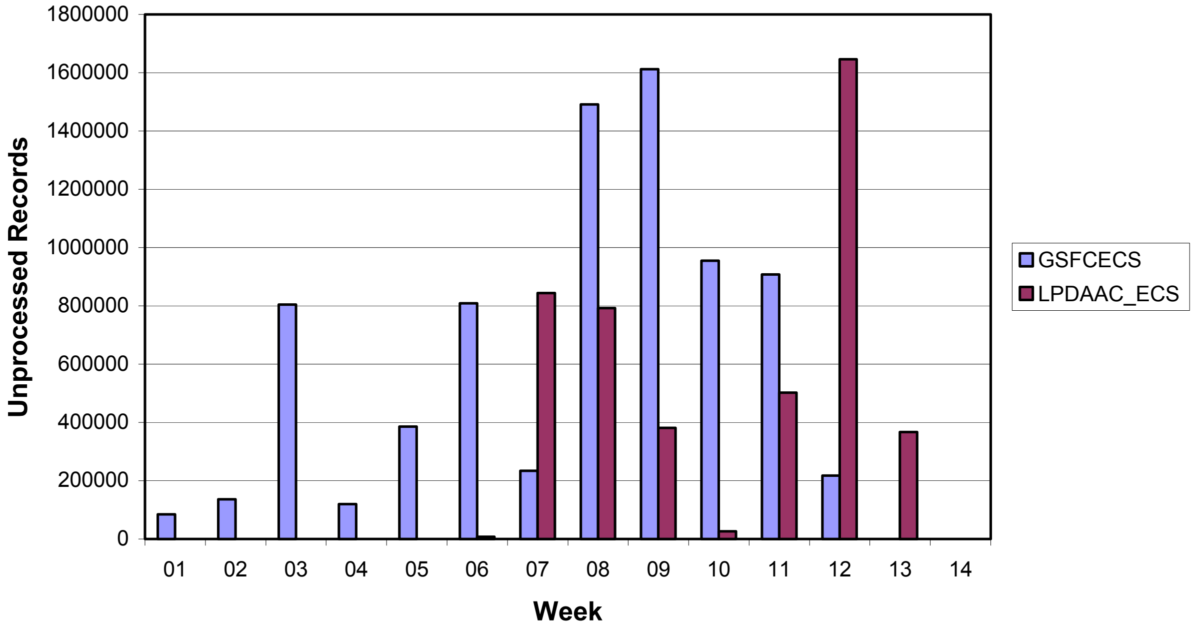

Figure 3. Records received but not processed in ECHO operational system by ProviderId by week.

4. ECHO Ingest Rates

At the start of each ingest "job", ECHO identifies all of the files available for the provider entity (e.g. LPDAAC_ECS) that is next in the ingest control file list. ECHO performs a pre-processing step that includes dividing the metadata ingest files into processing "chunks" of 1000 records or less. For each ingest job processed, ECHO creates detailed and summary log files that contain information on the ingest process, including start and end date/time, metadata files processed in the job, and details related to the transactions performed (e.g. number of inserts, replacements, deletions) for each chunk.

To extract metrics on ingest transactions, we parsed the ingest log files to extract relevant parameters (XML tags) and aggregated data for the transactions reported by chunk. The extracted information was stored in a database table (LogFile) that contains records from all log files created during the period. From this database, we were able to derive ingest rates and analyze characteristics of the associated ingest files that may influence ingest rates.

Figure 4 provides a summary of granule metadata ingest rates by week. *[Please note: Data points for GSFCECS ingest during weeks 01 and 02 need to be inserted in Figure 4]*. The overall average ingest rates observed for the GSFCECS, LPDAAC_ECS, and NSIDC_TS1 providers are 10.0, 11.5, and 18.6K records per hour, respectively.

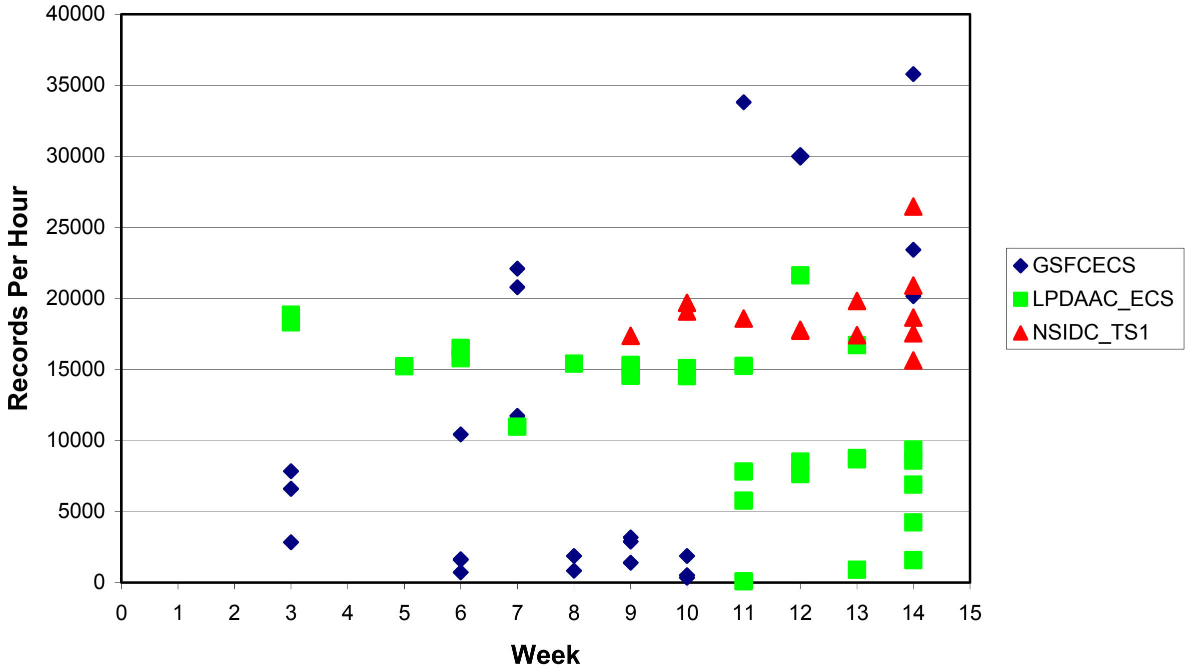

Figure 4. Ingest rates observed in ECHO operational system by week.

Browse metadata processing during the review period was significantly constrained by several key system hardware and software issues. As a result, only 8 ingest jobs for the period were identified to contain a sufficient number records (browse fraction > 0.20) for characterizing browse metadata ingest rates. All of the 8 ingest jobs considered belong to the AST_1B collection associated with the LPDAAC_ECS provider entity. The total number of records in the jobs considered was 336,876, and the range in number of browse records per job of 4.8K to 109.4K. For these ingest jobs, the minimum, maximum, and average observed rates of ingest were respectively, 3047, 6912, and 3982 records per hour.

Because the average ingest rate for browse metadata is substantially lower than that for granule metadata, the browse fraction in ingest jobs with mixed record types is a significant factor in over all ingest rate.

Another factor that substantially impacts ingest job performance is the fraction of granule replacements. **Figure 5** provides a summary of ingest rates by fraction of metadata replacement (granule update) transactions for ingest jobs where the fraction replaced was greater than 0.10. While there were several NSIDC_TS1 and GSFCECS ingest jobs that did not exhibit reduced job performance, the majority of the jobs analyzed with replacement fractions greater than 10% were severely impacted. For jobs containing only replacement transactions (Fraction Replaced = 1.0), the median ingest rate was 1158 records per hour, and there were 4 jobs with rates less than 800 records per hour.

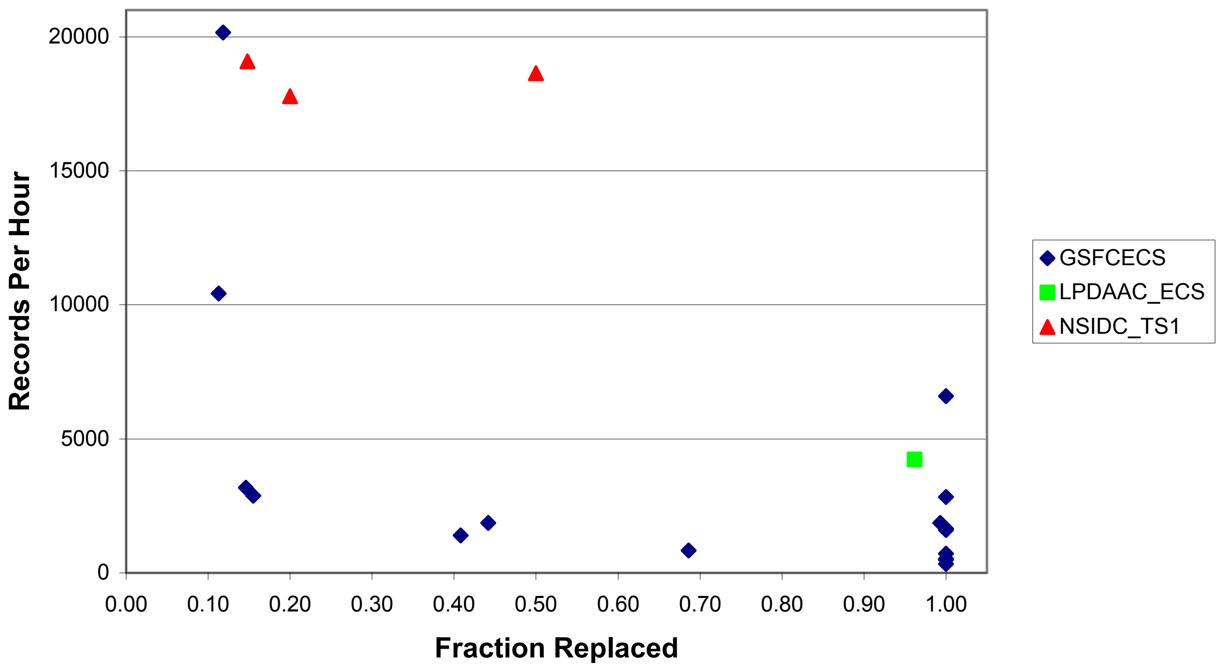

Figure 5. Ingest rates observed for granule replacement (update) transactions in operations during the quarter.

Because granule deletion transactions were not being processed correctly until ECHO version 5.0.1 ingest (after 12/18/2003), we did not attempt to analyze the affect of fraction deleted on ingest performance. However, early results from version 5.0.1 ingest indicate that granule deletion transactions may require substantially less resources during ingest than both insertions and replacements. Therefore, we are currently expecting to see a moderate increase in ingest rate for ingest jobs with large numbers of deletion transactions.

In analyzing ingest rates observed during the period, a number of other factors thought to have a possible impact on ingest rates were considered, including the number of records in database (provider schema) at time of ingest, the number of total records contained in the ingest job, and the density of metadata records (or lines per metadata record) in the input files. None of these factors were found to exhibit a statistically discernable influence on ingest rate. However, we believe that a combination of the lower metadata density and the smaller provider schema of the NSIDC_TS1 provider entity may explain the higher ingest rates observed for the granule metadata files (insertion transactions) generated by BMGT for this provider (~18,300 records per hour average for weeks 09-14), as compared to similar BMGT files generated for LPDAAC_ECS during the same period (~12,600 records/per hour average).

Part II: Current Projections

1. *Estimate of ECS DAAC Forward Processing and Historical Load*

In order to estimate the forward processing and historical load requirements for the ECS DAAC data providers, statistics on granules archived by the DAACs were collected using the Ad Hoc Archive Query function of the ESDIS Data Gathering and Reporting System (EDGRS) website (http://edgrs.gsfc.nasa.gov:8000/soo/aspdb_provider/edgrs3.asp).

In estimating forward processing requirements for the ECS DAACs, EDGRS granules archived data for the 12 calendar months in 2003 were compared to information collected in ECHO operations during the period under review. Since the EDGRS data for archived granules appear to be at least 15-20% high, we adjusted our current estimate of holdings accordingly.

Table 3 provides a summary of the estimated ECS DAAC monthly forward processing loads for granule metadata.

Provider	EDGRS Minimum	EDGRS Maximum	EDGRS Average	Adjusted Average (80% EDGRS Avg.)
GSFCECS	435,635	1,395,640	1,051,599	**841,279**
LPDAAC_ECS	501,815	802,817	638,384	**510,707**
NSIDC_ECS	98,483	223,019	165,823	**132,658**
LARC_ECS	97,110	340,427	210,621	**168,497**
ECS DAAC Forward Processing Total				**1,653,141**

Table 3. Estimates of ECS DAAC forward processing load for granule metadata per month.

To estimate the forward processing load for browse metadata, we used fractions of 0.16, 0.19, 0.10, and 0.15 applied to the adjusted average EDGRS data for granule records for the GSFCECS, LPDAAC_ECS, NSIDC_TS1, and LARC_ECS providers respectively. These fractions are derived from the ratio of forward processing browse to granule records observed for the three active ECS DAAC providers during the period. For LARC_ECS, we assume the same ratio as observed for LPDAAC_ECS. This results in a figure of approximately 270K records per month (63.0K per week) in forward processing load for browse metadata from the ECS DAACs.

[Please note: We are not all confident in the forward processing browse estimates. We are only using them temporarily as placeholders to enable the projection presented in the following section. All of the load estimates provided in this draft report will be updated as soon as we have input from the DAACs.]

For estimated historical load, EDGRS data for cumulative granules archived from 02/24/2000 through several end points in 2003 (March, August, November) were compared to information collected in ECHO operations during the period under review plus some limited additional information provided by the DAACs. Since the EDGRS data for archived granules appear to be at least 15-20% high, we adjusted our current estimate of holdings accordingly.

Table 4 provides a summary of the estimated ECS DAAC historical loads for granule metadata, the amount received to date by ECHO, the remaining granule records needed by ECHO and the remaining granule load to process.

Since the LP DAAC provided granule counts in their historical load export plan, we have at least one reasonable comparison point for our adjusted EDGRS estimate. For the historical load that LP DAAC plans to send to ECHO through 01/29/04, they estimated a total of 10,064,752 granules. Slightly more than 1.9 million granules were marked as having associated browse. A definitive comparison with the adjusted EDGRS figure cannot be made, since the LP DAAC plan did not specify the date through which their estimates were made. However, in noting that the numbers were fairly similar, the adjusted EDGRS value being only 6% higher, we feel comfortable using the adjusted EDGRS figures to make preliminary projections.

ECS Provider	EDGRS 80% Adj. Holdings Cum. thru Nov '03	Received by ECHO	Needed by ECHO	Current Backlog	Remaining to Process
GSFCECS	13,271,030	8,606,342	4,664,688 (35%)	5,564,801	10,229,489
LPDAAC_ECS	10,669,173	6,934,693	3,734,480 (35%)	4,546,937	8,281,417
NSIDC_ECS	2,677,549	0	2,677,549 (100%)	0	2,677,549
LARC_ECS	4,077,062	2,982,945	1,094,117 (27%)	0	4,077,062
Total	30,694,814	18,523,980	**12,170,834 (40%)**	10,111,738	**25,265,517**

Table 4. ECS DAAC historical load for granule metadata with status of ECHO acquisition and processing.

Using the same method described above for forward processing load, we estimate the browse records for the portion of the ECS DAAC historical load not yet received by ECHO to be around 1.9 million records. Adding this figure to the current GSFCECS and LPDAAC_ECS browse backlog provides a total estimate of approximately 4.2 million for the browse historical load remaining to be processed.

2. Projections for Completing Catalog of ECS DAAC Metadata

This section provides projections for the processing time required to populate the ECHO catalog with the complete historical load of the ECS DAACS and the time required to maintain currency with their forward processing. This is a top priority for ECHO Ops, and its completion will take precedence over other proposed activities this year. The information provided below, derived predominantly from the adjusted EDGRS statistics, suggests that we can meet our goal of completing the ECS DAAC historical load during FY2004.

Using the sum of the EDGRS adjusted monthly averages of forward processing load (granules) for the 4 ECS DAACs listed in Table 3, the provider ingest rates reported in *Part I: Section 4*, and the assumption of forward processing load for browse stated in the previous section, we estimate a weekly forward processing load that will require 39 processing hours to keep ECHO current with existing collection granules and browse.

Using the provider historical load information for granules listed in Table 4, we estimate that processing the current backlog of 10.1 million GSFCECS and LPDAAC_ECS granules will require between 620 to 1575 hours, depending on the fraction of replacement transactions (up to 1575 hours for 25% replacements). The 12.2 million new ECS DAAC granules and the nearly 3 million unprocessed LARC granules will require approximately 900 processing hours (0% replacement transactions assumed). In addition, the estimated 4.2 million browse records remaining in the historical load will require 1050 hours for processing. Thus the total time required for completing the historical load for the 4 ECS DAACS is on the order of 2500 to 3500 hours.

Processing time available per week (90% utilization):	151 hours
Time required for forward processing:	39 hours
Time remaining for historical load processing:	112 hours

Minimum historical load (2500 hrs.) completed in:	23 weeks
Maximum historical load (3500 hrs.) completed in:	32 weeks

Earliest completion of historical load:	late May 2004
Latest completion of historical load:	mid July 2004

Important Note to ECS DAAC Data Partners:

In distributing a draft of this report to the DAACs for review, we hope that we will be able to identify any important errors in our analysis early and adjust our plans as needed. We appreciate your cooperation in providing feedback as soon as possible. Please send comments and your own estimates of granule and browse metadata records to echo@killians.gsfc.nasa.gov or Jackie Kendall (301-867-2026; jackie_kendall@ssaihq.com).

3. Capacity for New Partners, New Missions, and Reprocessing

In addition to keeping up with forward processing of metadata files from current Partners, ECHO Ops is exploring the current system's capacity for ingesting metadata holdings from new data partners and new missions. ECHO Ops has spoken with several prospective data partners and is in the process of collecting their metadata holdings statistics.

ECHO Ops is also working on determining the occurrence of major reprocessing campaigns from current partners (e.g. Chris Lynnes' desire to refresh GSFCECS quarterly) and their affect on the system's ingest processes.

Table 5 provides a list of known prospective data partners and new missions.

Prospective Partner or New Mission	Datasets	Est. Records
Aura	OMI, HIRDLS, and MLS (GSFC) TES (LARC)	TBD
ASF V0 DAAC	TBD	TBD
JPL V0 DAAC	TBD	TBD
MQABI	MODIS Ocean QA Browse Imagery	TBD
NASA SSC	Science Data Purchase: AstroVision – 3 Collections EarthSat – 3 Collections EarthWatch – 2 Collections Positive Systems, Inc. – 1 Collection Space Imaging – 3 Collections	TBD
NSIDC V0 DAAC	TBD	TBD
SEDAC	TBD	TBD
USGS	NOAA AVHRR	TBD
Total records anticipated for new datasets		TBD

Table 5. Known prospective data partners and new missions with datasets proposed for near-future ECHO ingest.